DOCUMENT RESUME

ED 394 999 TM 024 995

AUTHOR Brandwein, Ann Cohen; Strawderman, William E.
TITLE James-Stein Estimation. Program Statistics Research, Technical Report No. 89-86.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-20
PUB DATE Apr 89
NOTE 46p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Equations (Mathematics); *Estimation (Mathematics); *Maximum Likelihood Statistics; *Statistical Distributions
IDENTIFIERS *James Stein Estimation; Nonnormal Distributions; *Shrinkage

ABSTRACT

This paper presents an expository development of James-Stein estimation with substantial emphasis on exact results for nonnormal location models. The themes of the paper are: (1) the improvement possible over the best invariant estimator via shrinkage estimation is not surprising but expected from a variety of perspectives; (2) the amount of shrinkage allowable to preserve domination over the best invariant estimator is, when properly interpreted, relatively free from the assumption of normality; and (3) the potential savings in risk are substantial when accompanied by good quality prior information. Relatively, much less emphasis is placed on choosing a particular shrinkage estimator than on demonstrating that shrinkage should produce worthwhile gains in problems where the error distribution is spherically symmetric. In addition, such gains are relatively robust with respect to assumptions concerning distribution and loss. (Contains 1 figure and 53 references.) (Author/SLD)

James-Stein Estimation

Ann Cohen Brandwein
and
William E. Strawderman

ETS

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 89-86

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

THE JAMES-STEIN ESTIMATION

Ann Cohen Brandwein

and

William E. Strawderman

Program Statistics Research
Technical Report No. 89-86

Research Report No. 89-20

Educational Testing Service
Princeton, New Jersey 08541-0001

April 1989

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

TABLE OF CONTENTS

Abstract

This paper presents an expository development of James–Stein estimation with substantial emphasis on exact results for non–normal location models.

The themes of the paper are a) that the improvement possible over the best invariant estimator via shrinkage estimation is not surprising but expected from a variety of perspectives; b) that the amount of shrinkage allowable to preserve domination over the best invariant estimator, is, when properly interpreted, relatively free from the assumption of normality; and c) the potential savings in risk are substantial when accompanied by good quality prior information.

1. Introduction.

This paper presents an expository development of James–Stein estimation with substantial emphasis on exact results for non–normal location models.

The themes of the paper are a) that the improvement possible over the best invariant estimator via shrinkage estimation is not surprising but expected from a variety of perspectives; b) that the amount of shrinkage allowable to preserve domination over the best invariant estimator, is, when properly interpreted, relatively free from the assumption of normality; and c) the potential savings in risk are substantial when accompanied by good quality prior information.

Relatively, much less emphasis is placed on choosing a particular shrinkage estimator than on demonstrating that shrinkage should produce worthwhile gains in problems where the error distribution is spherically symmetric. Additionally such gains are relatively robust with respect to assumptions concerning distribution and loss.

The basic problem, of course, is the estimation of the mean vector θ of a p–variate location parameter family. In the normal case (with identity covariance) for $p = 1$, the usual estimator, the sample mean, is the maximum likelihood estimator, the UMVUE, the best equivariant and minimax estimator for nearly arbitrary symmetric loss, and is admissible for essentially arbitrary symmetric loss. Admissibility for quadratic loss was first proved by Hodges and Lehmann (1950) and Girschick and Savage (1951) using the Cramer–Rao inequality and by Blyth (1951) using a limit of Bayes type argument.

For $p = 2$, the above properties also hold in the normal case. Stein (1956) proved admissibility using an information inequality argument. In that same paper however, Stein proved a result that astonished many and which has led to an enormous and rich literature of substantial importance in statistical theory and practice.

Stein (1956) showed that estimators of the form $(1 - a/(b+\|X\|^2))X$ dominate X for

a sufficiently small and b sufficiently large when $p \geq 3$. James and Stein (1961) sharpened the result and gave an explicit class of dominating estimators, $(1 - a/\|X\|^2)X$ for $0 < a < 2(p-2)$. They also indicated that a version of the result holds for general location equivariant estimators with finite fourth moment and for loss functions which are concave functions of squared error loss. Brown (1966) showed that inadmissibility of the best equivariant estimator of location holds for virtually all problems for $p \geq 3$, and, in Brown (1965), that admissibility tends to hold for $p = 2$. Minimaxity for all p follows from Kiefer (1957).

Section 2 gives a geometrical argument due to Stein which indicates that shrinkage might be expected to work under quite broad distributional assumptions.

Section 3 gives an empirical Bayes argument in the normal case which results in the usual James–Stein estimator.

Section 4 presents Stein's "unbiased estimator of risk" in the normal case and develops the basic theory for the standard James–Stein estimator in the normal case.

Section 5 describes a variety of extensions of the basic theory to cover shrinkage towards subspaces, Bayes minimax estimation, non–spherical shrinkage, and limited translation rules.

Section 6 considers extensions of the results of sections 4 and 5 to scale mixtures of normal distributions.

Section 7 presents generalizations to general spherically symmetric families of distributions, while section 8 indicates the applicability of earlier results to the multiple observation case.

Section 9 is concerned with results for non–spherical quadratic loss and for non–quadratic loss.

Section 10 presents some additional comments.

2. A Geometric Hint

As in much of the development of the subject, the following rough geometric argument is basically due to Stein (1962). Consider an observation vector X in p dimensions with mean vector θ and independent (or uncorrelated) components. Assume that the components have equal variance, σ^2. The situation is depicted in figure 1.

Figure 1 about here.

Since $E(X-\theta)' \theta = 0$ we expect $X-\theta$ and θ to be nearly orthogonal, especially for large $\|\theta\|$. Since $E\|X\|^2 = p\sigma^2 + \|\theta\|^2$, it appears that X as an estimator of θ might be too long, and that the projection of θ on X or something close to it might be a better estimator. This projection of course depends on θ and therefore isn't a valid estimator, but perhaps we can estimate it. If we denote this projection by (1–a)X the problem is to approximate a.

One way to do this is to assume that the angle between θ and $X-\theta$ is exactly a right

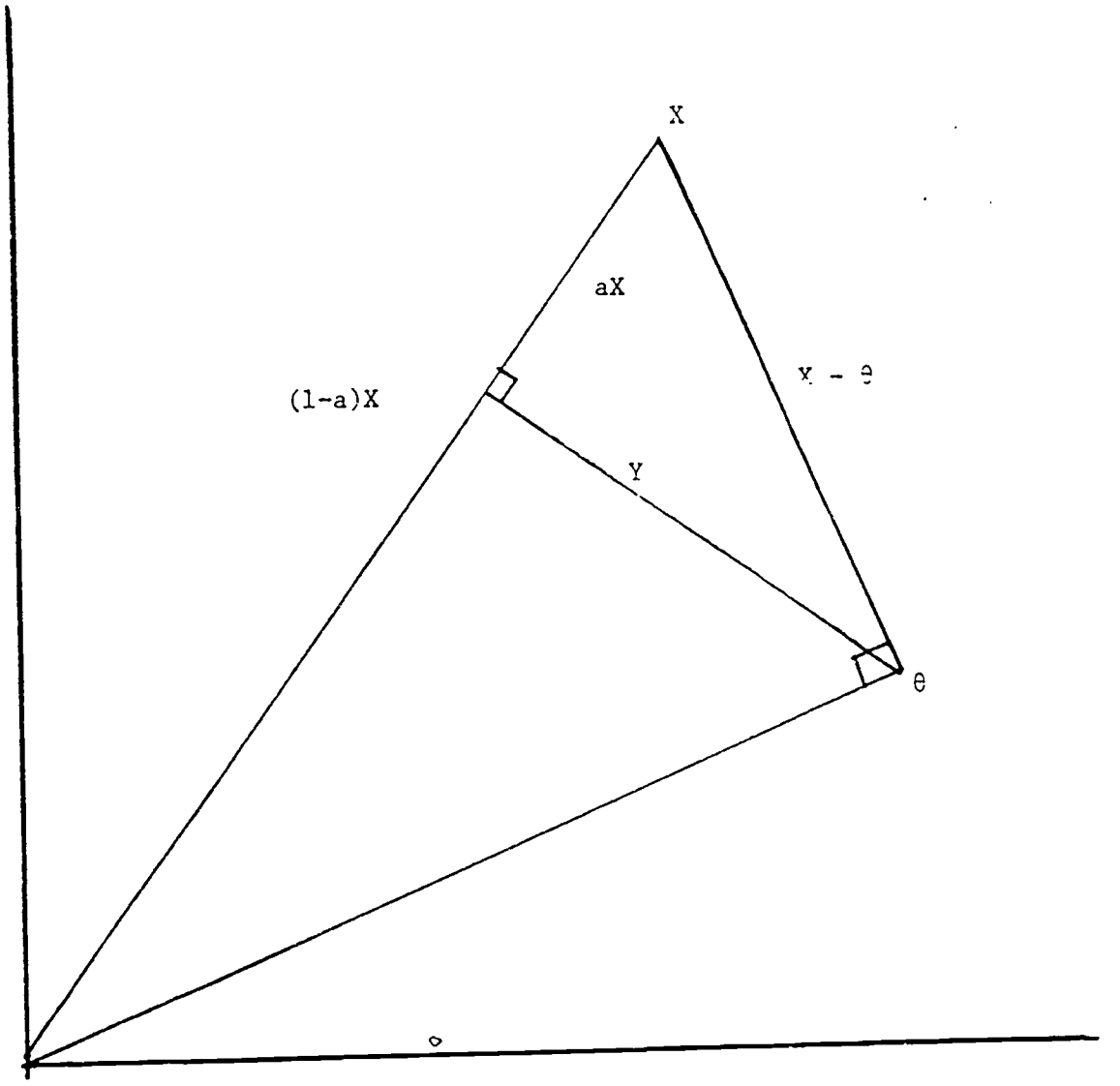

FIGURE 1

angle and to assume that $\|X\|^2$ is exactly equal to its expected value $\theta'\theta + p\sigma^2$ and similarly $\|X-\theta\|^2$ is equal to $p\sigma^2$. In this case we have $\|Y\|^2 = \|X-\theta\|^2 - \hat{a}^2\|X\|^2$ $= p\sigma^2 - \hat{a}^2\|X\|^2$ from triangle $\overline{BCD}$ and $\|Y\|^2 = \|\theta\|^2 - (1-\hat{a})^2\|X\|^2$ $= \|X\|^2 - p\sigma^2 - (1-\hat{a})^2\|X\|^2$ from triangle $\overline{ABD}$. Equating these expressions we obtain $p\sigma^2 - \hat{a}^2\|X\|^2 = \|X\|^2 - p\sigma^2 - (1-\hat{a})^2\|X\|^2$ or $(1-2\hat{a})\|X\|^2 = \|X\|^2 - 2p\sigma^2$. This gives $\hat{a} = p\sigma^2/\|X\|^2$ and the suggested estimator is $(1-\hat{a})\,X = (1 - \frac{p\sigma^2}{\|X\|^2})X$.

The above development does not particularly depend on normality of X or even that θ is a location vector. Unfortunately, it fails to be a proof of the inadmissibility of X, and also fails to distinguish between different values of p. It is however suggestive that the possibility of improving on the unbiased vector X by shrinkage toward the origin may be quite general.

3. An Empirical Bayes Argument

The following well known Empirical Bayes argument also leads to the James–Stein estimator. The origins of this argument, which we only briefly sketch, is unknown (to us). It has appeared numerous times in print (e.g. Lehmann (1983) p. 299).

Let X have a p–variate normal distribution with mean vector θ and (for simplicity) covariance matrix equal to σ^2 (known) times the identity. Suppose the prior distribution of θ is normal with mean vector 0 and covariance matrix equal to b times the identity,

here b is an unknown scalar. The posterior mean (assuming for the moment that b is known) is $(\frac{b}{\sigma^2+b})X = (1 - \frac{\sigma^2}{\sigma^2+b})X$.

One way to estimate the unknown scalar b is the following. Since $X-\theta$, conditional on θ is normal with mean vector 0 and covariance $\sigma^2 I$, $X-\theta$ and θ are independent. Hence $X = (X-\theta) + \theta$ is marginally distributed as a p–variate normal with mean vector 0 and covariance $(\sigma^2 + b)I$. Therefore, $\|X\|^2/(b + \sigma^2)$ has a central chi–square distribution with p degrees of freedom. It follows that marginally, $E(p-2)/\|X\|^2 = \frac{1}{\sigma^2+b}$ and hence $(1 - \frac{(p-2)\sigma^2}{\|X\|^2})X$ may reasonably be considered an Empirical Bayes estimator for the above normal prior with unknown scale. This estimator of course is exactly the usual James–Stein estimator.

4. Some Spherical Normal Theory

Let X have a p–variate normal distribution with mean vector θ and covariance matrix equal to the identity. The problem is to estimate θ with loss equal to

$$(4.1) \qquad L(\theta,\delta) = \|\delta - \theta\|^2 = \sum_{i=1}^{p} (\delta_i - \theta_i)^2.$$

Stein (1956) showed for $p \geq 3$, that the usual estimator $\delta_0(X) = X$ is dominated by $\delta_{a,b}(X) = (1 - a(b+\|X\|^2)^{-1})X$ provided a is sufficiently small and b is sufficiently large.

James and Stein (1961) showed that

$$(4.2) \qquad \delta_a(X) = (1 - a\|X\|^{-2})X$$

dominates X for $0 < a < 2(p-2)$ and that $a = p-2$ gives the uniformly best estimator in the class.

Their proof used the Poisson representation of the non–central chi–square distribution, but since the mid 1970's the "unbiased estimation of risk" technique of Stein (1981) been used and simplifies proofs substantially.

The technique, which we describe below, depends on the following Lemma.

LEMMA 4.1: Let $Y \sim N(\theta,1)$, then $E[h(Y)(Y-\theta)] = Cov(Y,h(Y)) = Eh'(Y)$ (provided e.g. that $h(Y)$ is the indefinite integral of $h'(Y)$, $\lim_{Y\to\pm\infty} h(Y)\exp[-\frac{1}{2}(Y-0)^2] = 0$ and all integrals are finite).

Proof: Integration by parts gives

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h(Y)(Y-\theta)\exp-[\tfrac{1}{2}(Y-\theta)^2]dy =$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h(Y)\frac{d}{dY}(-\exp(-\tfrac{1}{2}(Y-\theta)^2)dy$$

$$= -\frac{1}{\sqrt{2\pi}} h(Y)(\exp[-\tfrac{1}{2}(Y-\theta)^2)|_{-\infty}^{\infty}$$

$$+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h'(Y)\exp[-\tfrac{1}{2}(Y-\theta)^2]dy$$

$$= Eh'(Y)$$

□

To obtain an unbiased estimator of the risk of $\delta_a(X)$, write

$$(4.3) \qquad R(\theta,\delta_a(X)) = E\|(1-\frac{a}{\|X\|^2})X-\theta\|^2$$

$$= E\|X-\theta\|^2 + a^2E\frac{1}{\|X\|^2} - 2aE\frac{X'(X-\theta)}{\|X\|^2}$$

$$= p + a^2E\frac{1}{\|X\|^2} - 2a\sum_{i=1}^{p} E(\frac{X_i(X_i-\theta_i)}{\sum_{j=1}^{p} X_j^2})$$

$$= p + a^2E\frac{1}{\|X\|^2} - 2a\sum_{i=1}^{p} E(\frac{d}{dx_i}(\frac{X_i}{\Sigma X_j^2})) \quad \text{(by the lemma)}$$

$$= p + a^2E\frac{1}{\|X\|^2} - 2a\sum_{i=1}^{p} E\frac{\Sigma X_i^2 - 2X_i^2}{(\Sigma X_j^2)^2}$$

$$= p + a^2E\frac{1}{\|X\|^2} - 2aE\frac{[p\|X\|^2 - 2\|X\|^2]}{(\|X\|^2)^2}$$

$$= p + [a^2 - 2a(p-2)]\,E\frac{1}{\|X\|^2}\,.$$

Note that the quadratic $a^2 - 2a(p-2)$ is negative in the range $(0,2(p-2))$ and attains its minimum at $a = p-2$. Hence, we have the following result.

THEOREM 4.1 a. The estimator $\delta_a(X)$ in (4.2) dominates X for $0 < a < 2(p-2)$

(for $p \geq 3$). The estimator $\delta_{p-2}(X) = (1 - \frac{p-2}{X'X})X$ has the uniformly smallest risk of any estimator in the class.

b. The risk of $\delta_{p-2}(X)$ at $\theta = 0$ is equal to 2 for every dimension $p \geq 3$.

The proof of part b follows by noting that for $\theta = 0$, $\|X\|^2$ has a chi–square distribution with p degrees of freedom and hence by (4.3)

$$R(0,\delta_{p-2}(X)) = p - (p-2)^2 E_0 \frac{1}{\|X\|^2} = p - \frac{(p-2)^2}{p-2} = 2.$$

Part b suggests, particularly for large values of p, that very large savings in risk are possible over the classical estimator in the region near $\theta = 0$ at no cost of increased risk elsewhere.

Note also, that while substantial savings in risk are possible, the James–Stein estimator is itself inadmissible due to its strange behavior for small $X'X$. The shrinkage factor $(1 - \frac{p-2}{X'X})$ becomes negative for $X'X < p-2$. A better estimator is given by the "postive–part" estimator $(1 - \frac{p-2}{X'X})_+ X$. Interestingly, this estimator is itself inadmissible because it fails to be generalized Bayes, but to the authors' knowledge, no improved estimator has been found.

5. More Normal Theory

In the previous section we showed that the James–Stein estimator dominates the classical estimator in the (identity covariance) spherical normal case and that its risk at

the origin is equal to 2 regardless of the dimension of the problem. One may interpret this result as saying that if your prior guess that the origin is the true value is correct you may save substantially, but if you are wrong you loose nothing. This suggests, and it is obviously true, that a similar result holds for any arbitrary origin θ_0. That is, the estimator

$$\theta_0 + (1 - \frac{p-2}{\|X-\theta_0\|^2})(X - \theta_0) \tag{5.1}$$

will dominate the usual estimator and have risk equal to 2 at θ_0 provided $p \geq 3$.

Suppose it is believed that θ lies in V, an s dimensional subspace of R^p. Then letting the projection of X onto V be V and W = X–V, the projection of X onto the ortho–complement of V, we have the following result.

THEOREM 5.1 The estimator [(5.2) $V + (1 - \frac{(p-s-2)}{W'W})W$]dominates X, and has risk equal to s + 2 for all θ in V, provided $p - s \geq 3$. This is perhaps most easily seen by considering a canonical version, when V represents the first s coordinates and W the remaining p–s coordinates. The estimator (5.2) then uses the classical estimate on V and the James–Stein estimate on the (p–s) dimensional subspace $W = R^p - V$. In general the result follows by noting that V and $(1 - \frac{(p-s-2)}{W'W})W$ are independent (and orthogonal) and that the estimation problem breaks up into two orthogonal components ($\theta = \nu + \omega$, $\nu \varepsilon V \quad \omega \varepsilon W$).

One particularly important application of this idea is the estimator proposed by

Lindley (in discussion of Stein (1962)) where $V = \{\theta: \theta_1 = \theta_2 = ... = \theta_p\}$, dim V = 1, and the estimator (5.2) becomes

$$(5.3) \qquad \bar{X}1 + (1 - \frac{(p-3)}{E(X_i - \bar{X})^2})(X - \bar{X}1)$$

with $1 = (1,...,1)$.

As the preceding discussion and section 2 indicate there is a strong Bayesian connection to be made. In particular we indicated in section 2 that the James–Stein estimator could be viewed as an Empirical Bayes estimator for a normal prior with mean 0 and covariance matrix $\sigma^2 I$, with σ^2 unknown and estimated for the data. Strawderman (1971) established a more formal Bayesian (as opposed to Empirical Bayesian) connection along these lines. We now describe this development.

First, an extension of the basic James–Stein result due to Baranchick (1964, 1970) is helpful.

LEMMA 5.1: The estimator $(1 - \frac{r(X'X)}{X'X})X$ is minimax for the loss (4.1) provided $0 \leq r(\cdot) \leq 2(p-2)$, and $r(\cdot)$ is monotone increasing.

Proof: The proof in the case $\frac{r(X'X)}{X'X}$ satisfies the conditions of lemma 4.1 essentially follows that of Theorem 4.1. By Lemma 4.1

$$E(X-\theta)'X \frac{r(X'X)}{X'X} = (p-2)\, E \frac{r(X'X)}{X'X} + 2E \frac{r'(X'X)}{X'X} \geq (p-2)E \frac{r(X'X)}{X'X}$$

Hence

$$E\|(1 - \frac{r(X'X)}{X'X})X - \theta\|^2 = p + E\frac{r^2(X'X)}{X'X} - 2E\frac{(X-\theta)'X}{X'X}r(X'X)$$

$$\leq p + [2(p-2) - 2(p-2)]E\frac{r(X'X)}{X'X} = p$$

This lemma allows smoother shrinkage factors than the positive part James–Stein estimators and opens the possibility that generalized Bayes and perhaps proper–Bayes estimators may be found in the class (other than X itself). To this end, consider a two stage prior for θ such that at the first stage $\theta|\lambda \sim N(0, \frac{1-\lambda}{\lambda} I)$, and at the second stage $\lambda \sim (1-a)\lambda^{-a}$ (for $a < 1$). Then the Bayes estimator is given by

$$(5.5) \qquad E(\theta|X) = E[(E\theta|X,\lambda)|X]$$

$$= E[(1 - \frac{1}{1+(\frac{1-\lambda}{\lambda})})X|X] = [1 - E[\lambda|X]]X$$

A straightforward calculation gives

$$(5.6) \qquad E(\lambda|X) = \frac{1}{X'X}[p+2-2a - \frac{2\exp(-\frac{1}{2}X'X)}{\int_0^1 \lambda^{\frac{1}{2}p-a}\exp(-\frac{\lambda}{2}X'X)d\lambda}] \cdot$$

$$= \frac{r(X'X)}{X'X}$$

Where $r(X'X)$ is defined to be the term in brackets on the right side of (5.5). Since $r(X'X) \leq p+2-2a$, and since $\int_0^1 \lambda^{\frac{1}{2}p-a}\exp\{-\frac{X'X}{2}(1-\lambda)\}d\lambda$ is increasing, the conditions of

Lemma 5.1 will be satisfied provided $p+2-2a \leq 2(p-2)$ or equivalently

$$a \geq \frac{6-p}{2}. \tag{5.7}$$

Since in order for λ^{-a} (the second stage prior) to be integrable we must have $a < 1$, it is seen that (5.7) is satisfied for such an a provided that $p \geq 5$. Hence we have

THEOREM 5.2. a) For $p \geq 5$ the proper–Bayes estimator (5.5) is minimax provided $a \geq \frac{1}{2}(6-p)$.

b) For $p \geq 3$ the estimator (5.5) is generalized Bayes and minimax provided $\frac{1}{2}(6-p) \leq a < \frac{1}{2}(p+2)$.

The proof of b) follows by noting that (5.6) makes sense (i.e. the generalized Bayes estimator exists) provided $\frac{1}{2}\,p-a > -1$, which is equivalent to the right inequality. Note that the double inequality holds only if $p > 2$. Strawderman (1972) showed that no proper Bayes minimax estimators exist for $p < 5$.

We briefly take a broader view and describe a result of Stein (1981) concerning minimaxity of (generalized) Bayes estimators. If $\pi(\theta)$ is the (generalized) prior density, then the (generalized) Bayes estimator is given by

$$\delta_\pi(X) = \frac{\int \theta\, e^{-\frac{1}{2}\|X-\theta\|^2} \pi(\theta)d\theta}{\int e^{-\frac{1}{2}\|X-\theta\|^2} \pi(\theta)d\theta} \tag{5.6}$$

$$= X + \nabla \log \int e^{-\frac{1}{2}\|X-\theta\|^2} \pi(\theta)d\theta$$

$$= X + \nabla \log f(X) = X + \frac{\nabla f(X)}{f(X)}$$

where $\nabla = (\frac{\partial}{\partial X_1}, \frac{\partial}{\partial X_2},\ldots,\frac{\partial}{\partial X_p})$, and $f(X)$ is (essentially) the posterior density of θ given X.

An easy application of Lemma 4.1 gives the following very general unbiased estimate of risk for a nearly arbitrary estimator of the form $X+g(x)$.

LEMMA 5.2 Let $\delta(X) = X+g(X)$ be such that $g(\cdot)$ is almost differentiable and such that $\sum_{i=1}^{p} E|\nabla_i g_i(X)| < \infty$. Then

$$E\|X+g(X) - \theta\|^2 = p + E_\theta[\|g(X)\|^2 + 2\nabla \circ g(X)].$$

Hence if $\|g(X)\|^2 + 2\nabla \circ g(X) < 0$ for all X then $X + g(X)$ dominates X.

(Note that Theorem 4.1 and Lemma 5.1 are special cases).

Application of the lemma to a (generalized) Bayes estimator of the form (5.6) gives

$$(5.7) \qquad R(\theta,\delta_\pi) = E\|X + \frac{\nabla f(X)}{f(X)} - \theta\|^2$$

$$= p + E[\frac{\|\nabla f(X)\|^2}{f^2(X)} + 2\,\frac{f(X)\nabla^2 f(X) - \|\nabla f(X)\|^2}{f^2(X)}]$$

$$= p + E[\frac{\nabla^2 f(X)}{f(X)} - \frac{\|\nabla f(X)\|^2}{f^2(X)}]$$

$$\leq p + E\,\frac{\nabla^2 f(X)}{f(X)}$$

(Actually the penultimate expression can be simplified to equal $p + 4E\,\frac{\nabla^2\sqrt{f(X)}}{\sqrt{f(X)}}$ where $\nabla^2 f(X) = \sum_{i=1}^{p} \frac{\partial^2}{\partial^2 X_i} f(X)$, the Laplacian of $f(\cdot)$).

A function $f(X)$ such that $\nabla^2 f(X) \leq 0\ \forall\ X$ is called superharmonic. It has the property that the average of the function over a sphere of radius r about a point X_0 is never greater than $f(X_0)$ for all X_0 and $r > 0$. Further it is easily seen that convex combinations of superharmonic functions are superharmonic. It follows then that if $\pi(\theta)$ is a superharmonic prior, then $f(x) = \int \frac{e^{-\frac{1}{2}\|X-\theta\|^2}}{(\sqrt{2\pi})^p}\,\pi(\theta)d\theta$ is also superharmonic. We then have the neat result.

THEOREM 5.2. If $\pi(\theta)$ is superharmonic,then the estimator (5.6) is minimax.

Incidentally, note that if $\pi_\gamma(\theta)$ is superharmonic for each γ, then so is $\pi^*(\theta) = \int \pi_\gamma(\theta)\, dF(\gamma)$ for any distribution $F(\cdot)$. This opens up a nice class of multiple shrinkage minimax estimators due to George (1985).

To conclude this section we present three examples which illustrate the utility of lemma 5.2.

EXAMPLE 5.1 (Stein 1981)

Let

$$\delta(X) = X - \frac{\beta AX}{X'BX} \tag{5.8}$$

Then

$$E\|\delta(X) - \theta\|^2 = p + E[\beta^2 \frac{X'A^2X}{(X'BX)^2} - 2\nabla o \frac{\beta AX}{X'BX}]$$

$$= p + E[\beta^2 \frac{X'A^2X}{(XBX)^2} - \frac{2\beta t\, r\, A}{X'BX} + \frac{4\beta X'ABX}{(X'BX)^2}]$$

(provided the expectations exist).

If A is a fixed symmetric matrix, $B = [(\text{tr } A)I - 2A]^{-1}A^2$, and $2A < (\text{tr } A)I$ (the largest eigenvalue of A is less than $\frac{1}{2}$ the trace of A). Then

$$E\|\delta(X) - \theta\|^2 = p + (\beta^2 - 2\beta)E\frac{X'A^2X}{(X'BX)^2} .$$

Hence δ dominates X provided $0 < \beta < 2$. Furthermore $\beta = 1$ is the uniformly best choice.

Stein (1981) gives an intereting application of this result to three term symmetric moving averages of the form

$$\hat{\theta}_i = X_i - \lambda(X)(X_i - \tfrac{1}{2}(X_{i-1} + X_{i+1})) \tag{5.9}$$

where $X_0 = X_p$. Here

$$A_{ij} = \begin{cases} -\frac{1}{2} & \text{if } j - i \equiv \pm 1 \pmod{p} \\ 1 & \text{if } j - i \equiv 0 \pmod{p} \\ 0 & \text{otherwise} \end{cases}$$

The characteristic roots are $1 - \cos(2\pi \frac{j}{p}) \leq 2$ with $-[\frac{p}{2}] \leq j < [\frac{p}{2}]$ giving $\text{tr}(A) = p$. Hence for $p \geq 5$ (5.9) dominates X for $\lambda(x) = \frac{AX}{X'BX}$ as in (5.8).

EXAMPLE 5.2 (Berger 1980).

Let the generalized prior density be given by

$$\int_0^1 [\det \beta(\lambda)]^{-\frac{1}{2}} \exp[-\frac{(\theta'-\mu)\beta^{-1}(\lambda)(\theta-\mu)}{2}] \lambda^{n-1-\frac{p}{2}} d\lambda$$

where $\beta(\lambda) = \lambda^{-1}$ C–I for $0 < \lambda < 1$, C–I positive definite and $n > 0$. Here the distribution of θ given λ is $N(\mu,(B(\lambda))$ and reduces to the two stage prior of Strawderman discussed earlier in this section if C = I. The (generalized) Bayes estimator is given by

(5.10) $$\delta_n(X) = \mu + (I - \frac{r_n((X-\mu)'C^{-1}(x-\mu))C^{-1}}{(X-\mu)'C^{-1}(X-\mu)})(X-\mu)$$

where $$r_n(V) = \frac{V\int_0^1 \lambda^n \exp(-(\frac{\lambda V}{2}) d\lambda}{\int_0^1 \lambda^{n-1}\exp(-\frac{\lambda V}{2}) d\lambda}$$

$$= 2n(1 - [n\int_0^1 \exp(\frac{(\lambda-1)V}{2} d\lambda]^{-1})$$

It follows from Lemma 5.2 (as in Example 5.1) that if $(2+n)\ Ch_{max}C^{-1} \leq tr\ C^{-1}$ then $\delta_n(X)$ is minimax.

EXAMPLE 5.3 (Stein 1981).

Efron and Morris (1971, 1972) considered estimators which modified the James–Stein estimator by requiring that no coordinate moves by more than a preassigned quantity C. Stein gave an alternative "limited translation" rule based on order statistics as follows. Let $Z_i = |X_i|$ and $Z_{(1)} < Z_{(2)} < \ldots < Z_{(p)}$, be the order statistics. Fix K a positive integer (a large fraction of p) and consider $\delta(X) = X + g(X)$ when

$$g_i(x) = \begin{cases} -\dfrac{a}{\Sigma(X_j^2 \wedge Z_{(K)}^2)} & X_i \text{ if } |X_i| \leq Z_{(K)} \\ -\dfrac{a}{\Sigma(X_j^2 \wedge Z_{(K)}^2} & Z_{(K)} \operatorname{sgn} X_i \text{ if } |X_i| > Z_{(K)} \end{cases}$$

where $a \wedge b = \min(a,b)$.

Application of Lemma 5.2 gives

$$E\|\delta_{(X)} - \theta\|^2 = p + [a^2 - 2(K-2)a]\, E\left[\frac{1}{\Sigma(X_j^2 \wedge Z_{(K)}^2}\right]$$

Hence the estimator is minimax if $0 < a < 2(K-2)$ and the uniformly best choice of a is $K-2$.

6. Scale Mixtures Of Normals

Stein (1956) showed that the usual estimator of a location vector could be improved upon quite generally for $p \geq 3$ and Brown (1966) stubstantially extended this conclusion to essentially arbitrary loss functions. Explicit results of the James–Stein type however were restricted to the case of the normal distribution. Strawderman (1974) considered scale mixtures of multivariate normal distributions as follows. Let X have density $f(\|X - \theta\|^2)$ where

$$(6.1) \qquad f(\|X - \theta\|^2) = (\sqrt{2\pi})^{-p} \int \exp\left[-\frac{1}{2\sigma^2}\|X - \theta\|^2\right]\sigma^{-p} dG(\sigma)$$

where $G(\cdot)$ is a known distribution function. The object is to estimate θ with loss (4.1).

Such a random variable X clearly has the interpretation that given σ, X is normal with mean vector θ and covariance matrix $\sigma^2 I$. The unconditional distribution of σ is

$G(\cdot)$. This interpretation together with Lemma 5.2 allows the following calculation of the risk of a smooth estimator $X + g(X)$.

$$
\begin{aligned}
(6.2) \qquad E\|X + g(X) - \theta\|^2 &= E^{\sigma}[E^{X|\sigma}\|X + g(X) - \theta\|^2 | \sigma^2] \\
&= E^{\sigma}\sigma^2[E^{X|\sigma}\|\frac{X}{\sigma} + \frac{g(\frac{X}{\sigma}\cdot\sigma)}{\sigma} - \frac{\theta}{\sigma}\|^2 | \sigma] \\
&= E^{\sigma}\sigma^2 E^{X|\sigma}[p + \frac{1}{\sigma^2}\|g(\frac{X}{\sigma}\cdot\sigma)\|^2 \\
&\qquad + 2\frac{1}{\sigma}\nabla_{\frac{X}{\sigma}}\cdot g(\frac{X}{\sigma}\cdot\sigma) | \sigma] \\
&= p\, E^{\sigma}\sigma^2 + E^{\sigma}E^{X|\sigma}[\|g(X)\|^2 + 2\sigma^2\nabla\cdot g(X) | \sigma]
\end{aligned}
$$

where $\nabla_{\frac{X}{\sigma}}\cdot g(\frac{X}{\sigma}\cdot\sigma) = \nabla\cdot g(u\cdot\sigma)|_{u=\frac{X}{\sigma}}$.

For estimators of the James–Stein type, $g(X) = -\frac{a}{X'X}$ and, $\nabla\cdot g(X) = -\frac{a(p-2)}{X'X}$ and hence

$$
\begin{aligned}
E\|(1 - \frac{a}{X'X})X - \theta\|^2 &= pE^{\sigma}\sigma^2 + E^{\sigma}E^{X|\sigma}[\frac{a^2}{X'X} - \frac{2a(p-2)}{X'X}\sigma^2 | \sigma^2] \\
&= pE^{\sigma}\sigma^2 + E^{\sigma}[(\frac{a^2}{\sigma^2} - 2a(p-2))E^{X|\sigma}(\frac{\sigma^2}{X'X} | \sigma)].
\end{aligned}
$$

Note that $X'X/\sigma^2$ given σ^2 is distributed as a non–central χ^2 with p degrees of

freedom and non–centrality parameter $\frac{\theta'\theta}{\sigma^2}$. Hence $E^{X|\sigma}(\frac{\sigma^2}{X'X}|\sigma)$ is an increasing function of σ^2. Since $\frac{a^2}{\sigma^2} - 2a(p-2)$ is decreasing in σ^2 we have

$$(6.3) \qquad E\|(1 - \frac{a}{X'X})X - \theta\|^2 \le pE^{\sigma}\sigma^2 + E^{\sigma}[\frac{a^2}{\sigma^2} - 2a(p-2)]E^{\sigma}E^{X|\sigma}[\frac{\sigma^2}{X'X}|\sigma] \le pE\sigma^2$$

provided $0 < a < 2(p-2)/E^{\sigma}\frac{1}{\sigma^2} = 2/E_0\frac{1}{X'X}$. Therefore we have the following result.

THEOREM 6.1. Let X have the distribution 6.1 for $p \ge 3$. Then the estimator $(1 - \frac{a}{X'X})X$ dominates X (for the loss 4.1) provided $0 < a < 2/E_0(\frac{1}{X'X})$.

It is interesting to note that this result reduces to Theorem 4.1 a if the distribution of σ is degenerate at $\sigma = 1$. Furthermore, the shrinkage factor $a = 2/E_0(\frac{1}{X'X})$ is an upper bound for any distribution such that each coordinate has mean 0, as an easy calculation shows. What is remarkable about Theorem 6.1 is that if the shrinkage factor is interpreted properly, the James–Stein result extends directly to the entire class of scale mixtures of normal distributions.

Note that this class includes (if $1/\sigma^2 \sim \chi^2_K$) the family of multivariate –t distributions with tails of the order $(1 + \theta'\theta)^{-(\frac{p+K}{2})}$ as well as the family of normal distributions.

The geometrical argument of section 2 which hinted at shrinkage factors of the order of $p\sigma^2$ regardless of normality is thus validated for a wide class of distributions.

Chou and Strawderman (1986) extended this result to include estimators of the type studied in Lemma 5.2. Here is a simple form of their result.

THEOREM 6.2. Let X be as in Theorem 6.1, and let g(X) be such that

a) $\|g(X)\|^2 + 2\nabla \cdot g(X) \leq 0$

b) $g(bX) = \frac{1}{b}g(X)$ ($g(\cdot)$ is homogeneous of degree -1)

c) $\{X: \|g(X)\|^2 > c\}$ is convex for each $c \geq 0$.

Then $X + ag(X)$ dominates X for $0 < a < 2/E(1/\sigma^2)$.

Proof: By (6.2)

$$E\|X + ag(X) - \theta\|^2 = pE\sigma^2 + E^{\sigma}E^{X|\sigma}[a^2\|g(X)\|^2 + 2a\sigma^2\nabla \cdot g(X)|\sigma]$$

$$\leq pE\sigma^2 + E^{\sigma}E^{X|\sigma}[\|g(X)\|^2[a^2 - 2a\sigma^2]|\sigma^2]$$

$$= pE\sigma^2 + E^{\sigma}E^{X|\sigma}[\|g(\tfrac{X}{\sigma})\|^2[\frac{a^2}{\sigma^2} - 2a]|\sigma^2]$$

$$\leq pE\sigma^2.$$

The last inequality follows since $E\|g(\frac{X}{\sigma})\|^2$ is increasing in σ (by Anderson's Theorem) and $[\frac{a^2}{\sigma^2} - 2a]$ is decreasing in σ^2, and $E[\frac{a}{\sigma^2} - 2a] < 0$ if $0 < a < \frac{2}{E(1/\sigma^2)}$.

Hence versions of the estimators of Example 5.1, 5.2 and 5.3 extend to the scale mixture of normal families.

7. Other Spherically Symmetric Distributions

Extensions of James–Stein type results to distributions other than scale mixtures of normal distributions are due to Berger (1975), Brandwein and Strawderman (1978), Brandwein (1979), and Bock (1985).

We present a new proof of Brandwein's result which gives shrinkage factor for $p \geq 4$ which holds uniformly for all spherically symmetric distributions such that $EX'X < \infty$ and $E(X'X)^{-1}$ is fixed. The factor given is the best possible and is attained for uniform distributions concentrated on a spherical "shell" of radius R.

The spirit of the proof is to first obtain the result for spherical shells and to extend it by use of a technical lemma to mixtures of such distributions. Since the class of spherically symmetric distributions is precisely those obtained by scale mixtures of spherical shells, the desired result follows.

Suppose X has a p–variate spherically symmetric density of the form $f(\|X-\theta\|^2)$. Then $X = \theta + U$ where U has density $f(\|U\|^2)$.

We will need the following facts:

F1: $$V = \left[\frac{\theta' U}{\|\theta\| \ \|U\|}\right]^2 = (\cos(\theta, U))^2$$

has a Beta $(\frac{1}{2}, \frac{p-1}{2})$ distribution independent of $\|U\|^2$ (see Dempster (1969) p.272).

F2: If p(v) is the density of V then

$$p_\gamma(v) = c(\gamma)p(v)[1 + \gamma - \tfrac{4\gamma v}{1+\gamma}]^{-1} \text{ and } p^*_\gamma(v) = c^*(\gamma)p(v)[1 + \gamma - \tfrac{4\gamma v}{1+\gamma}]^{-2}$$

have monotone likelihood ratio decreasing for $0 < \gamma < 1$ and increasing for $1 < \gamma < \infty$

(Proof: easy calculation).

We will prove

THEOREM 7.1: If X has density $f(\|X-\theta\|^2)$ and $E\|X\|^2$ and $E\|X\|^{-2}$ are finite then for $p \geq 4$, $(1-\frac{a}{X'X})X$ dominates X for $0 < a < 2(\frac{p-2}{p})[E_0\|X\|^{-2}]^{-1}$ for loss (4.1).

Proof: $R(\theta,(1-\frac{a}{X'X})X) =$

$$(7.1) \qquad E[\|(1 - \frac{a}{X'X})\, X - \theta\|^2] = E\|X-\theta\|^2 + a^2 E\,\frac{1}{X'X} - 2aE\frac{X'(X-\theta)}{X'X}.$$

Let $X = U + \theta$, $U'U = R^2$ and use the fact that the distributions of U and $-U$ coincide to get

$$(7.2) R(\theta, (1 - \frac{a}{X'X})X) = EU'U + E[\frac{a^2}{2}[\frac{1}{U'U + \theta'\theta + 2U'\theta} + \frac{1}{U'U + \theta'\theta - 2U'\theta}]$$

$$- \frac{2a}{2}[\frac{U'U + U'\theta}{U'U + \theta'\theta + 2U'\theta} + \frac{U'U - U'\theta}{U'U + \theta'\theta - 2U'\theta}]$$

$$= ER^2 + E\Big\{\frac{a^2(R^2 + \theta'\theta)}{(R^2 + \theta'\theta)^2 - 4(\theta'U)^2} - \frac{2a[R^2(R^2 + \theta'\theta) - 2(\theta'U)^2]}{(R^2 + \theta'\theta)^2 - 4(\theta'U)^2}\Big\}.$$

Since $R(\theta,X) = E\|X - \theta\|^2 = ER^2$, we have, letting $V = \frac{(\theta'U)^2}{\|\theta\|^2\|U\|^2}$ $(0 \leq V \leq 1)$

(7.3) $$R(\theta,(1-\frac{a}{X'X})X) - R(\theta,X)$$

$$= E\left\{\frac{a^2(R^2 + \theta'\theta) - 2a[R^2(R^2 + \theta'\theta) - 2\theta'\theta R^2 V]}{(R^2 + \theta'\theta)^2 - 4\theta'\theta R^2 V}\right\}$$

$$\leq EE\left\{\frac{a^2 - 2aR^2(1 - 2V)}{(R^2 + \theta'\theta) - 4\theta'\theta R^2 V[\theta'\theta + R^2]^{-1}} | R\right\}$$

$$= EE\left\{\frac{\frac{a^2}{R^2} - 2a(1-2V)}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}V[1 + \frac{\theta'\theta}{R^2}]^{-1}} | R\right\}.$$

Now using fact F2 that if $p(v)$ is the density of V then

$\frac{c(\frac{\theta'\theta}{R^2})p(v)}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}v[1 + \frac{\theta'\theta}{R^2}]^{-1}} = p_{\frac{\theta'\theta}{R^2}}(v)$ has monotone decreasing likelihood ratio if $\frac{\theta'\theta}{R^2} < 1$ and monotone increasing likelihood ratio if $\frac{\theta'\theta}{R^2} > 1$ to conclude that

(7.4) $$\int_0^1 v\, p_{\frac{\theta'\theta}{R^2}}(v)dv = \frac{E[\frac{V}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}V[1 + \frac{\theta'\theta}{R^2}]^{-1}} | R]}{E[\frac{1}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}V[1 + \frac{\theta'\theta}{R^2}]^{-1}} | R]} \times$$

$$\leq \max[\int_0^1 v p_0(v)dv, \lim_{\gamma\to\infty}\int_0^1 v p_\gamma(v)dv] = \frac{1}{p}.$$

The last equality follows since $p_\gamma(v) = P_{1/\gamma}(v)$ and hence

$\lim_{\gamma\to\infty} \int_0^1 vp_\gamma(v)dv = \lim_{\gamma\to\infty} \int_0^1 vp_{1/\gamma}(v)dv = \int_0^1 vp_0(v)dv = \frac{1}{p}$. Here we also use Fl that $p_0(v)$ is a Beta $(\frac{1}{2}, \frac{p-1}{2})$.

Hence combining (7.3) and (7.4)

$$(7.5) \qquad R(\theta,(1-\frac{a}{X'X})X) - R(\theta,X) \le EE\Big\{\frac{\frac{a^2}{R^2} - 2a(\frac{p-2}{p})}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}V[1 + \frac{\theta'\theta}{R^2}]^{-1}} \Big| R\Big\}.$$

We will show below that

$$E\Big[\frac{1}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}V[1 + \frac{\theta'\theta}{R^2}]^{-1}} \Big| R\Big]$$

is decreasing in $\frac{\theta'\theta}{R^2}$ and hence increasing in R^2 for fixed $\|\theta\|^2$. This together with the fact that $\frac{a^2}{R^2} - 2a(\frac{p-2}{p})$ is decreasing in R^2 implies that

$$R(\theta,(1-\frac{a}{X'X})X) - R(\theta,X) \le E[\frac{a^2}{R^2} - 2a(\frac{p-2}{p})]E\frac{1}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2}V[1 + \frac{\theta'\theta}{R^2}]^{-1}} < 0$$

provided $a^2 E \frac{1}{R^2} - 2(\frac{p-2}{p}) < 0$. The Theorem follows since $E_0 \frac{1}{\|X\|^2} = E\frac{1}{R^2}$.

It remains to show $E[\frac{1}{1 + \frac{\theta'\theta}{R^2} - 4\frac{\theta'\theta}{R^2} V[1 + \frac{\theta'\theta}{R^2}]^{-1}} | R]$ is decreasing in $\frac{\theta'\theta}{R^2}$ to complete the proof. Note first

$$(7.6) \quad \frac{d}{d\gamma}\int_0^1 \frac{p(v)dv}{1+\gamma - 4\gamma v[1+\gamma]^{-1}} = \frac{1}{(1+\gamma)^2}\int_0^1 \frac{4v - (\gamma+1)^2}{[1+\gamma - 4\gamma v[1+\gamma]^{-1}]^2} p(v)dv$$

Clearly if $\gamma \geq 1$ this derivative is negative. To prove (7.6) is negative in the range $0 < \gamma < 1$ use F2 and show

$$c^*(\gamma)\int_0^1 \frac{4v - (\gamma+1)^2}{[1+\gamma - 4\gamma v[1+\gamma]^{-1}]^2} p(v)dv \leq c^*(\gamma)\int_0^1 \frac{4v - 1}{[1+\gamma - 4\gamma v[1+\gamma]^{-1}]^2} p(v)dv$$

$$< \int_0^1 (4v-1)p(v)dv = (\frac{4}{p} - 1) \leq 0 \text{ if } p \geq 4 \text{ where } c^*(\gamma) = [\int_0^1 \frac{p(v)}{[1+\gamma - 4\gamma v[1+\gamma]^{-1}]^2} dv]^{-1}.$$

This completes the proof.

We have noted that the factor $2(\frac{p-2}{p})/E_0\|X\|^{-2}$ is the best possible constant which holds uniformly for all spherically symmetric distributions with $E_0\|X\|^{-2}$ fixed. For specific distributions, obviously, better results are possible (see Bock (1985)). It is

remarkable however, that the best possible constant for any distribution can be no larger than $2/E_0\|X\|^{-2}$ as can be easily seen by calculating the risk at 0. Hence the factor given in Theorem 7.1 which applies uniformly is surprisingly close to the best that can be attained for any given distribution.

We note for completeness that the results of Theorem 6.1 for mixtures of normals and Theorem 7.1 for spherically symmetric distributions can be extended to prove minimaxity of estimators of the form $(1 - \frac{ar(X'X)}{X'X})X$. The conditions on a are as in their respective theorems and the conditions on $r(\cdot)$ are : a) $0 < r(\cdot) < 1$; b) $r(Y)$ is monotone nondecreasing; and c) $r(Y)/Y$ is monotone nonincreasing.

8. Multiple Observations

So far we have concentrated our attention on improving the estimator X based on a single observation from a population with density $f(\|X-\theta\|^2)$. Suppose we have a sample $X_1,..,X_n$ from such a population and the problem is to estimate the p–dimensional vector θ with loss (4.1).

In this case the natural estimator is Pitman's estimator, one version of which is given by $\delta(X,Y) = X_1 - E_0[X_1|Y]$, where $Y = (X_1 - X_n, X_2 - X_n,...,X_{n-1} - X_n)$. This estimator which is minimax and best among equivariant estimators is inadmissible if $p \geq 3$. For $n > 2$, $\overline{X}$, the vector of sample averages, is Pitman's estimator if and only if the

population is normal. For non–normal populations, Pitman's estimator is typically difficult to calculate and its distribution tends to be analytically intractable. A variety of other estimators such as $\overline{X}$, or the M.L.E. might be used instead, all of which are in the class of estimators equivariant under the orthogonal and location groups. That is for any such estimator

$$\delta(QX_1 - c, QX_2 - c,...,QX_n - c) = Q\,\delta(X_1,...,X_n) - c.$$

Since the sampling distribution of any such estimator (when sampling from a spherically symmetric distribution) is itself spherically symmetric, Theorem 7.1 applies and we may conclude that $\delta(X_1,...,X_n)$ is dominated by $(1 - \frac{a}{\|\delta(X_1,...,X_n)\|^2})\,\delta(X_1,...,X_n)$ for $0 < a < [2(p-2)]/[pE_0\|\delta(X_1,...,X_n)\|^{-2}]$.

9. Other Loss Functions

There are two major lines of development relating to generalizations concerning the loss function (4.1). The first is to consider general quadratic loss given by

(9.1) $$L(\delta,\theta) = (\delta-\theta)'D(\delta-\theta)$$

where D is a given pxp positive definite matrix. The second relates to non–quadratic loss.

The earliest results for loss 9.1 in the normal case are due to Bhattacharya (1966). An early representative result is the following due to Bock (1975). Let

$$\delta(X) = [I - aD^{-\frac{1}{2}}CD^{\frac{1}{2}}\|X\|^{-2}]X \tag{9.2}$$

where C is a known positive definite matrix. Let ξ_L be the largest eigenvalue of $D^{\frac{1}{2}}CD^{\frac{1}{2}}$ and γ_L be the largest eigenvalue of $D^{\frac{1}{2}}C^2D^{\frac{1}{2}}$.

THEOREM 9.1. Let $X \sim N_p(\theta,I)$, then the estimator (9.2) dominates X under loss (9.1) provided $0 < a < 2[\mathrm{tr}CD - 2\xi_L]/\gamma_L$.

Proof:

$$R(\theta,X) - R(\theta,\delta) = a^2E[\frac{X'D^{\frac{1}{2}}C^2D^{\frac{1}{2}}X}{(X'X)^2}] - 2aE[\frac{X'D^{\frac{1}{2}}C\ D^{\frac{1}{2}}(X-\theta)}{(X'X)}]$$

$$= a^2E\frac{X'D^{\frac{1}{2}}C^2D^{\frac{1}{2}}X}{(X'X)^2} - 2a\nabla\cdot\frac{D^{\frac{1}{2}}C\ D^{\frac{1}{2}}X}{(X'X)}$$

(by lemma 4.1 as in example 5.1)

$$= a^2E\frac{X'D^{\frac{1}{2}}C^2D^{\frac{1}{2}}X}{(X'X)^2}] - 2a\ \frac{[X'X\ \mathrm{tr}D^{\frac{1}{2}}C\ D^{\frac{1}{2}} - 2X'D^{\frac{1}{2}}CD^{\frac{1}{2}}X]}{(X'X)^2}$$

$$< E \frac{1}{X'X}[a^2\gamma_L - 2a(\text{tr } CD - 2\xi_L)]$$

$$< 0 .$$

There are a number of results of this type for estimators of the form

(9.3) $$[I - \frac{a\, r(X)\, B}{X'CX}]X$$ which may be proved in much the same way.

It is worth noting that the problem of estimating the mean vector θ when $X \sim N(\theta,\Sigma)$ with Σ known, loss given by (9.1), and estimators of the form (9.3) is essentially reducible to the case $\Sigma = I$. In this more general setting a variety of justifications for different choices of B and C in (9.3) have been given from the robust Bayesian perspective (Berger (1982)) from the ridge regression perspective (Thisted (1976), Strawderman (1978), Draper and Van Nostrand (1979), Casella (1980)), and from an empirical Bayesian perspective (Efron and Morris (1973, 1975), Morris (1983)) among others.

A variety of results covering non–normal situations have been found by Berger (1975) and Chou and Strawderman (1986) in the scale mixture of normal case, and by Brandwein and Strawderman (1978) in the spherically symmetric unimodal case, and by

Brandwein (1979) in the spherically symmetric case.

In particular, Brandwein's (1979) result for X from a spherically symmetric distribution replaces the upper bound for a in Theorem 9.1 by

$$\frac{2}{p}(\mathrm{tr}cD - 2\xi_L)\gamma_L^{-1}[E_0(X'X)^{-1}]^{-1}$$

In the normal case this is (p–2)/p times the upper bound in Theorem 9.1 and again, the degree of shrinkage allowed is relatively unaffected by the assumption of normality.

Results concerning extensions to non–quadratic loss are relatively few. Berger (1978) has results in the normal case for polynomial loss. Brandwein and Strawderman (1980) and Bock (1985) have results for losses of the form

$$L(\theta,\delta) = f(\|\delta-\theta\|^2) \tag{9.4}$$

where $f(\cdot)$ is an increasing concave function. Here is a version of Brandwein and Strawderman's result.

THEOREM 9.2. Let X have a spherically symmetric distribution with $p \geq 4$. Then $\delta(X) = (1 - \frac{a}{X'X})X$ dominates X for the loss (9.4) provided $0 < E_G f'(R^2) < \infty$ and $0 < [2(p-2)]/[pE_H R^{-2}]$ where $G(\cdot)$ is the cdf of $R = \|X-\theta\|$ and $H(R) = \int_0^R f'(s^2)dG(s)/\int_0^\infty f'(s^2)dG(s)$. E_G and E_H denote expected values under the cdf

$G(\cdot)$ and $H(\cdot)$ respectively.

Proof: We use the fact that $f(\cdot)$ concave implies

$f(\|X-\theta\|^2 + u) \leq f(\|X-\theta\|^2) + uf'(\|X-\theta\|^2)$ to obtain

(9.5)
$$R(\theta,\delta) - R(\theta,X)$$
$$= E[f(\|X-\theta\|^2 + a^2/X'X - 2a(X-\theta)'X/X'X) - f(\|X-\theta\|^2)]$$
$$\leq E\{f'(\|X-\theta\|^2)[a^2/X'X - 2a(X-\theta)'X/X'X]\}$$
$$= EE[(a^2/X'X - 2a(X-\theta)'X/X'X)f'(R^2)\|X-\theta\| = R]$$

But it follows from Theorem 7.1 that the last expression in (9.5) is negative provided $0 < a < 2(p-2)/[pE_H R^{-2}]$. Hence the theorem follows.

The proof for estimators of the form $(1 - \frac{ar(X'X)}{X'X})X$; where $r(\cdot)$ satisfies the conditions of the remarks following the proof of Theorem 7.1, is essentially identical to the above proof.

As an application of this result to the spherical normal case, let $X \sim N_p(\theta,I)$, $L(\theta,\delta) = \|\delta-\theta\|^q$, $0 < q < 2$. Hence $f(u) = u^{q/2}$, $f'(u) = q/2\ u^{(q-2)/2}$ and since $R^2 = \|X-\theta\|^2 \sim X^2_p$, $E_H R^{-2} = E_G R^{-2+q-2}/E_G R^{q-2} = (p+q-4)^{-1}$. Hence we are assured that the estimator $(1 - a/X'X)X$ dominates X for $0 < a < 2(p-2)(1 - (4-q)/p)$, for $p \geq 4$.

The most important results for non–quadratic loss are those for confidence set estimation. Hwang and Casella (1982) showed that the usual spherical confidence set centered at X may be dominated by one centered at an appropriate (positive–part) James–Stein estimator. Hwang and Chen (1986) extend domination results for confidence sets centered at positive–part James–Stein estimators to non–normal settings.

10. Comments

1. Unknown Scale. We have assumed throughout that the scale is known. For the $N_p(\theta,\sigma^2 I)$ distribution, if an estimator of σ^2 is available which is distributed as a multiple of a chi–square distribution independently of $\overline{X}$ this case causes no difficulty. The original James–Stein paper treats this problem as do several others. In the non–normal case much less is known. One can make some progress if an independent estimate of the scale is known (at least in the mixture of normal case (see Bravo and MacGibbon (1987)) but such an assumption seems unwarranted generally.

2. Non Spherically Symmetric Distributions. The discussion of section 9 can be extended to handle distributions of the form $f((X-\theta)'\Sigma^{-1}(X-\theta))$ where Σ is a known positive–definite matrix by working with the random vector $\Sigma^{-\frac{1}{2}}X$ which has a spherically symmetric distribution. In cases where the whole problem is not spherically symmetric a tension

between "being Bayes" (doing well on the average) and being minimax (never doing worse than the best invariant estimator) often develops. It typically happens that minimax estimators will shrink coordinates with larger variances relatively less than will Bayes estimators. The phenomenon is complicated by the fact that for quadratic loss, the minimax estimator will depend on the choice of D in (9.1) while the Bayes estimator will not. See Berger (1985) and references therein for more details. The current recommendations for choice of shrinkage procedures in such situations seems to favor a Bayesian or Empirical Bayesian basis as opposed to a purely minimax one even among more classically oriented decision theorists. This seems to be at least partly on the grounds that minimaxity may be too strict a requirement here, and that relaxation to something like ε–minimaxity might preserve the large gains possible (near the origin, say) at a slight cost for "large" values of θ in certain directions.

3. Independent Coordinates. It can be argued that a much more natural class of problems than the ones we have been considering are those non–normal location problems where the coordinates are independent. Since sphericity and independence implies normality we have, unfortunately described no results for the non–normal case. Shinozaki (1984), Miceli and Strawderman (1986,1988) have some results for independent non–normal observations but the results are not nearly as extensive as for the spherical case.

4. Applications. We have said little about applications of James–Stein estimation. Efron and Morris in a series of papers (1971,1972, 1975, 1976, 1977 and others) fostered the application of shrinkage estimation and addressed a number of practical considerations including the unequal variance case, shrinkage in groups, and limited translation estimators. Most of the published applications have had an empirical Bayes orientation. For some examples the reader is referred to Efron and Morris (1973,1975), Casella (1985), Green and Strawderman (1985,1986) and Braun et al (1983).

REFERENCES

Baranchik, A.J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Stanford University. Technical Report No. 5i.

Baranchik, A.J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.* 41, 642–645.

Berger, J. (1975). Minimax estimation of location vectors for a wide class of densities. *Ann. Statist.* 3, 1318–1328.

Berger, J. (1978). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* 4, 223–226.

Berger, J. (1978). Minimax estimation of a multivariate normal mean under polynomial loss. *J. Multivariate Anal.* 8, 173–180.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* 8, 716–761.

Berger, J. (1982). Bayesian robustness and the Stein effect. *J. Amer. Statist. Assoc.* 77, 358–368.

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis, Second Edition Springer–Verlag, New York.

Bhattacharya, P.K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math. Statist.* 37, 1819–1824.

Blyth, C.R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* 22, 22–42.

Bock, M.E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* 3, 209–218.

Bock, M.E. (1985). Minimax estimators that shift towards a hypersphere for location vectors of spherically symmetric distributions. *J. Multivariate Anal.* 17, 127–147.

Brandwein, A.C. and Strawderman, W.E. (1978). Minimax estimation of location parameters for spherically symmetric unimodal distributions. *Ann. Statist.* 6, 377–416.

Brandwein, A.C. (1979). Minimax estimation of the mean of spherically symmetric distributions under general quadratic loss. *J. Multivariate Anal.* 9, 579–588.

Brandwein, A.C. and Strawderman, W.E. (1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *Ann. Statist.* 8, 279–284.

Braun, H., Jones, D., Rubin, D. and Thayer, D. (1983). Empirical Bayes estimation of coefficients in the general lineal model from data of deficient rank. *Psychometrika*, 48, 171–181.

Bravo, G. and MacGibbon, B. (1987). Improved shrinkage estimators for the mean vector of a scale mixture of normals with unknown variance. To appear in the *Canadian Journal of Statistics.*

Brown, L.D. (1965). On the admissibility of invariant estimators of two dimensional location parameters. Mimeo. Notes. Birbeck College, London.

Brown, L.D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37**, 1087–1135.

Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.* 8, 1036–1056.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39, 83–87.

Chou, J.P. and Strawderman, W.E. (1986). Minimax estimation of means of multivariate normal mixtures. Technical Report, Rutgers University.

Dempster, A.P. (1969). Elements of Continuous Multivariate Analysis. Addison–Wesley Publishing Company, Inc.

Draper, N.R. and Van Nostrand, R.C. (1979). Ridge regression and James–Stein estimation: review and comments. *Technometrics* 21, 451–466.

Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and Empirical Bayes estimators – Part I: The Bayes case. *J. Amer. Statist. Assoc.* **66**, 807–815.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and Empirical Bayes estimators – Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67**, 130–139.

Efron, B and Morris, C. (1973a). Stein's estimation rule and its competitors–an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

Efron, B. and Morris, C. (1973b). Combining possibly related estimation problems. *J. Roy Statist. Soc.* (Ser B) **35**, 379–421.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311–331.

Efron, B. and Morris, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* 4, 22–32.

Efron, B. and Morris, C. (1977). Comment. *J. Amer. Statist. Assoc.* **72**, 91–93.

George, E.I. (1985). Shrinkage towards multiple points and subspaces. Technical Report, Graduate School of Business, University of Chicago, Chicago.

Girschick, M.A. and Savage, L.J. (1951). Bayes and minimax estimates for quadratic loss functions. Proc. Second Berkeley Symp. Math. Prob. 1, 53–73. University of California Press Berkeley.

Green, E. and Strawderman, W.E. (1985). The use of Bayes/Empirical Bayes estimation in individual volume equation development. Forest Science 31, 975–990.

Green, E. and Strawderman, W.E. (1986). Stein rule estimation of coefficients for 18 eastern hardwood cubic volume equations. *Canadian Journal of Forest Research* 16, 249–255.

Hodges, J.L and Lehmann, E.L. (1950). Some problems in minimax point estimation. *Ann. Math. Statist.* 21, 187–197.

Hwang, J.T. and Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.* 10, 868–881.

Hwang, J.T. and Chen, J. (1986). Improved confidence sets for the coefficients of a linear model with spherically symmetric errors. *Ann. Statist.* 14, 44–460.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In Proc. Fourth Berkeley Symp. Math. Statist. Probab. 1, 361–380. University of California Press, Berkeley.

Kiefer, J. (1957). Invariance, minimax sequential estimation and continuous time processes. *Ann. Math. Statist.* 28, 537–601.

Lehmann, E.L. (1983). Theory of Point Estimation. Wiley, New York.

Miceli, R. and Strawderman, W.E. (1986). Minimax estimation for certain independent component distributions under weighted squared error loss. *Communications in Statistics*, 15, 2191–2200.

Miceli, R. and Strawderman, W.E. (1988). "Almost Arbitrary" Estimates of location for independent component variance mixtures of normal variates. To appear in Prob. and Statist. Letters.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* 78, 47–65.

Shinozaki, N. (1984). Simultaneous estimation of location parameters under quadratic loss. *Ann. Statist.* 12, 322–335.

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In Proc. Third Berkeley Symp. Math. Statist. Probab. 1, 197–206. University of California Press, Berkeley.

Stein, C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc.* (Ser. B) 24, 265–296.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *J. Roy Statist.* 9, 1135–1151.

Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* 42, 385–388.

Strawderman, W.E. (1972). On the existence of proper Bayes minimax estimators of the mean of a multivariate normal distribution. In Proc. Sixth Berkeley Symp. Math. Statist. Probab. 1, 51–55. University of California Press, Berkeley.

Strawderman, W.E. (1974). Minimax estimation of location parameters for certain spherically symmetric distributions. *J. Multivariate Anal.* 4, 255–264.

Strawderman, W.E. (1978). Minimax adaptive generalized ridge regression estimators. *J. Amer. Statist. Assoc.* 73, 623–627.

Thisted, R.A. (1976). Ridge regression minimax estimation, and empirical Bayes methods. Stanford University, Department of Statistics, Technical Report 28, December.